A New Randomized Method for Solving Large Linear Systems

Elnur Gasanov[1] Vladislav Elsukov[2] Peter Richtárik[1,2]

King Abdullah University of Science and Technology [1] Moscow Institute of Physics and Technology [2]

The Problem

Given $\mathbf{A} \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, solve $\mathbf{A}x = b$. That is, find $x \in \mathcal{L} \overset{\text{def}}{=} \{x \mid \mathbf{A}x = b\}$.

- **Challenge:** This problem is difficult when $m \gg n$ (e.g., $m = 10^9$)
- **Goal:** Design a new randomize method for finding an approximate solution quickly

The "Basic Method" (BM) [1]

One of the ways to solve the problem is via the following algorithm:

Input: Matrix $\mathbf{A}$, vector b
Parameters: $x_0 \in \mathbb{R}^n$, stepsize $\omega > 0$, positive definite matrix $\mathbf{B}$, distribution $\mathcal{D}$ from which to sample matrices
for $k = 0, 1, 2, \dots$ **do**
  Draw a fresh sample $\mathbf{S}_k \sim \mathcal{D}$
  $\mathbf{H}_k^{\text{BM}} = \mathbf{S}_k(\mathbf{S}_k^\top \mathbf{A}\mathbf{B}^{-1}\mathbf{A}^\top \mathbf{S}_k)^\dagger \mathbf{S}_k^\top$
  $x_{k+1} = x_k - \omega \mathbf{B}^{-1}\mathbf{A}^\top \mathbf{H}_k^{\text{BM}}(\mathbf{A}x_k - b)$
end
Output: $x_k \approx x_* \overset{\text{def}}{=} \arg\min_{x:\mathbf{A}x=b} \|x - x_0\|_{\mathbf{B}}$

Remarks:

- BM generalizes the randomized Kaczmarz method [2] (who only considered $\omega = 1$, $\mathbf{B} = \mathbf{I}$ and very special $\mathcal{D}$)
- BM also generalizes [3] (who only considered $\omega = 1$)
- The matrix $\mathbf{G}^{\text{BM}} \overset{\text{def}}{=} \mathbb{E}_{\mathbf{S}_k\sim\mathcal{D}}[\mathbf{H}_k^{\text{BM}}]$ controls the speed of the method

A New Method (NM)

- We propose a new method for solving the problem.
- The method does not require require to calculate the Moore-Penrose pseudoinverse appearing in $\mathbf{H}_k^{\text{BM}}$.

Input: Matrix $\mathbf{A}$, vector b
Parameters: Same as BM; plus: carefully designed parameters $L_{\mathbf{S}} > 0$ associated with every matrix $\mathbf{S}$
for $k = 0, 1, 2, \dots$ **do**
  Draw a fresh sample $\mathbf{S}_k \sim \mathcal{D}$
  $\mathbf{H}_k^{\text{NM}} = \frac{1}{L_{\mathbf{S}_k}} \mathbf{S}_k \mathbf{S}_k^\top$
  $x_{k+1} = x_k - \omega \mathbf{B}^{-1}\mathbf{A}^\top \mathbf{H}_k^{\text{NM}}(\mathbf{A}x_k - b)$
end
Output: $x_k \approx x_*$

Remarks:

- $L_{\mathbf{S}_k}$ is required to satisfy
$$L_{\mathbf{S}} \geq \lambda_{\max}(\mathbf{S}^\top \mathbf{A}\mathbf{B}^{-1}\mathbf{A}^\top \mathbf{S})$$
for the method to work
- The matrix $\mathbf{G}^{\text{NM}} \overset{\text{def}}{=} \mathbb{E}_{\mathbf{S}_k\sim\mathcal{D}}[\mathbf{H}_k^{\text{NM}}]$ controls the speed of the method

NM vs SGD

Theorem [GER'19] NM is SGD applied to the problem
$$\min_{x\in\mathbb{R}^n} f(x) \overset{\text{def}}{=} \mathbb{E}\left[f_{\mathbf{S}}(x) \overset{\text{def}}{=} \tfrac{1}{2L_{\mathbf{S}}}\|\mathbf{A}x - b\|^2_{\mathbf{S}\mathbf{S}^\top}\right]$$
That is, NM is equivalent to the method:

1. Sample $\mathbf{S}_k \sim \mathcal{D}$
2. $x_{k+1} = x_k - \omega \nabla f_{\mathbf{S}_k}(x_k)$

Exactness

Define matrix: $\mathbf{\Omega} \overset{\text{def}}{=} \mathbf{B}^{-\frac{1}{2}}\mathbf{A}^\top \mathbf{G}^{\text{NM}}\mathbf{A}\mathbf{B}^{-\frac{1}{2}}$.

Theorem [GER'19] These statements are equivalent:

1. $\mathcal{L} = \mathcal{X} \overset{\text{def}}{=} \text{Argmin} f(x)$ ("exactness")
2. $\text{Null}\left((\mathbf{G}^{\text{NM}})^{\frac{1}{2}}\mathbf{A}\right) = \text{Null}(\mathbf{A})$
3. $\text{Null}(\mathbf{G}^{\text{NM}}) \cap \text{Range}(\mathbf{A}) = \emptyset$
4. $\text{Null}(\mathbf{\Omega}) = \text{Null}(\mathbf{A}\mathbf{B}^{-\frac{1}{2}})$

Moreover, if $\mathcal{D}$ is absolutely continuous, then exactness for the NM is equivalent to exactness for BM

Convergence of iterates

Theorem [GER'19] Let $\mathbf{\Omega} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$ be the eigenvalue decomposition. Then the random iterates generated by NM converge linearly as follows:
$$\|\mathbb{E}[x_k - x_*]\|^2_{\mathbf{B}} \leq \rho^k \|x_0 - x_*\|^2_{\mathbf{B}},$$
where $\rho = \max_{i:\lambda_i>0}(1 - \omega\lambda_i(\mathbf{\Omega}))^2$. Moreover,
$$\mathbb{E}\|x_k - x_*\|^2_{\mathbf{B}} \leq (1-\theta)^k \|x_0 - x_*\|^2_{\mathbf{B}},$$
where $\theta = \lambda^+_{\min}(\mathbf{\Omega})$.

Convergence of function values

Let $\lambda^+_{\min}$ and $\lambda_{\max}$ be the smallest positive and largest eigenvalues of $\mathbf{\Omega}$, respectively. Choose $\omega \in [0; \frac{2\lambda^+_{\min}}{\lambda_{\max}}]$. Then
$$\mathbb{E}[f(x_k)] \leq (1 - 2\lambda^+_{\min}\omega + \lambda_{\max}\omega^2)^k f(x_0) \quad (1)$$
The optimal rate is achieved for $\omega = \lambda^+_{\min}/\lambda_{\max}$, in which case we get the bound
$$\mathbb{E}[f(x_k)] \leq \left(1 - (\lambda^+_{\min})^2/\lambda_{\max}\right)^k f(x_0)$$

Comparison of rates

If exactness holds for both methods, then
$$\lambda^+_{\min}(\mathbf{\Omega}) \leq \lambda^+_{\min}(\mathbf{B}^{-\frac{1}{2}}\mathbf{A}^\top \mathbf{G}^{\text{BM}}\mathbf{A}\mathbf{B}^{-\frac{1}{2}})$$
If $\mathbf{B} = \mathbf{I}$, then
$$\frac{\lambda^+_{\min}(\mathbf{A}^\top \mathbf{G}_{\text{bm}}\mathbf{A})}{\lambda^+_{\min}(\mathbf{A}^\top \mathbf{G}_{\text{nm}}\mathbf{A})} \leq \sup_{\mathbf{S}} \frac{\sqrt{\lambda_{\max}(\mathbf{S}^\top \mathbf{A}\mathbf{A}^\top \mathbf{S})}}{\lambda^+_{\min}(\mathbf{S}^\top \mathbf{A}\mathbf{A}^\top \mathbf{S})}$$
Theorem [GER'19] Let $\mathbf{B} = \mathbf{I}$ and assume the ith row of $\mathbf{A}$ satisfies $\|\mathbf{A}_{i:}\|_2 = 1$ for all i. Further, let $\mathcal{D}$ be defined as follows: $\mathbf{S}_k$ is a random column submatrix of $\mathbf{I}$ consisting of 2 columns. Finally, assume that $|\langle \mathbf{A}_{i:}, \mathbf{A}_{j:}\rangle| \leq \alpha_0 < 1$ for all $i \neq j$. Then
$$\frac{\lambda^+_{\min}(\mathbf{A}^\top \mathbf{G}_{\text{BM}}\mathbf{A})}{\lambda^+_{\min}(\mathbf{A}^\top \mathbf{G}_{\text{NM}}\mathbf{A})} \leq \frac{\sqrt{1+\varepsilon}}{1-\varepsilon}.$$

This means that NM is at most $\frac{\sqrt{1+\varepsilon}}{1-\varepsilon}$ times slower than BM in terms of iterations. (However, it can be much faster in terms of cost of one iteration.)

Adaptive computation of $L_{\mathbf{S}}$

If in each iteration $L_{\mathbf{S}_k}$ satisfies
$$\|\mathbf{A}x_k - b\|^2_{\mathbf{H}_k^{\text{BM}}} \leq L_{\mathbf{S}_k}\|\mathbf{A}x_k - b\|^2_{\mathbf{H}_k^{\text{NM}}},$$
then for NM we have $\mathbb{E}\|x_k - x_*\|^2_{\mathbf{B}} \to 0$ linearly.

References

[1] Peter Richtárik and Martin Takáč.
Stochastic reformulations of linear systems: Algorithms and convergence theory.
arXiv preprint arXiv:1706.01108, 2017.

[2] Thomas Strohmer and Roman Vershynin.
A randomized Kaczmarz algorithm with exponential convergence.
Journal of Fourier Analysis and Applications, 2008.

[3] Robert M Gower and Peter Richtárik.
Randomized iterative methods for linear systems.
SIAM Journal on Matrix Analysis and Applications, 36(4):1660–1690, 2015.